ARMY RESEARCH LABORATORY

Visual Analytics for Exploration of a High-Dimensional Structure

by Andrew M. Neiderer

ARL-TN-532 **April 2013**

Approved for public release; distribution is unlimited.

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Aberdeen Proving Ground, MD 21005-5067

ARL-TN-532 **April 2013**

Visual Analytics for Exploration of a High-Dimensional Structure

Andrew M. Neiderer
Computational and Information Sciences Directorate, ARL

Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
April 2013	Final	September 2012–January 2013

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Visual Analytics for Exploration of a High-Dimensional Structure	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
Andrew M. Neiderer	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
U.S. Army Research Laboratory ATTN: RDRL-CII-C Aberdeen Proving Ground, MD 21005-5067	ARL-TN-532
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report is about a stage of the knowledge discovery in databases (KDD) process used to find possible patterns in high-dimensional data (HDD): data distributed in the form of a geometrical locus (or object) in HDD space or data close to some manifold. The emphasis is on data mining for exploratory data analytics of the HDD and dimensionality reduction by feature selection/extraction, which is necessary for a two- or three-dimensional representation of the HDD for exploratory visual analytics. Such a description allows us to navigate and interact with the data.

15. SUBJECT TERMS

visual analytics, dimensionality reduction, high-dimensional data, knowledge discovery in databases

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Andrew M. Neiderer
Unclassified	Unclassified	Unclassified	UU	18	19b. TELEPHONE NUMBER (*Include area code*) 410-278-3203

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

1. Introduction

The U.S. Army Research Laboratory is using a knowledge discovery in databases (KDD) approach to find patterns and structure, if any, in documentation (intelligence reports, news articles, etc.) concerning terrorist-related events (see figure 1). The intent is to expedite the KDD process for such activity so that it can be disrupted or thwarted. Sometimes there are indicators, but often the relevant information is buried within a massive amount of other data. High-dimensional data (HDD) may increase the chances of an incorrect pattern. This so-called "curse of dimensionality" may include anomalies in the raw data caused by (1) sensor malfunction in extreme environmental conditions or (2) errors resulting from computer program code, such as the floor function approximation. And then there is the challenge of possible multilingual data mining under a time constraint. All of these, and more, are reasons why anticipating a terrorist event is an extremely difficult task.

This report addresses the stage in the KDD process from dimensionality reduction (DR) to interpretation—namely, feature selection (FS), feature extraction (FE), and data-mining methods. The approach used here involves transforming unstructured text, such as that from the Global Terrorism Database (*1*), and is very HDD, to a two- or three-dimensional (2-D/3-D) data representation suitable for visual analytics (VA) application. Exploratory data analytics (EDA), which is closely related to the field of data mining, is used to discover knowledge in the data.

The next section describes how an EDA problem in HDD space becomes an exploratory visual analytics (EVA) one in 2-D/3-D Euclidean space—when a point set embedded in a high-dimensional geometric space is transformed to a visually based distribution shape or structure. To our knowledge, successive application of FS (section 2.1) prior to FE (section 2.2) has not been considered, and thus the usefulness is still being evaluated. FS is semantic-preserving, while FE destroys semantics but allows us to examine the underlying relationships in the data even though the meaning of the variables is lost. The assumption here is that humans most effectively understand HDD as 2-D/3-D objects/structure in Euclidean space (*2*).

Section 3 discusses EVA, which allows for 3-D geometric manipulation of the data. For a 2-D view, affine transformation(s) is/are followed by an orthographic projection onto an arbitrary plane. Sometimes just looking at the data from different views reveals something interesting or informative; otherwise, analyzing the same unstructured text would be difficult.

Finally, we conclude and suggest where future efforts can be made for a more effective terrorist KDD.

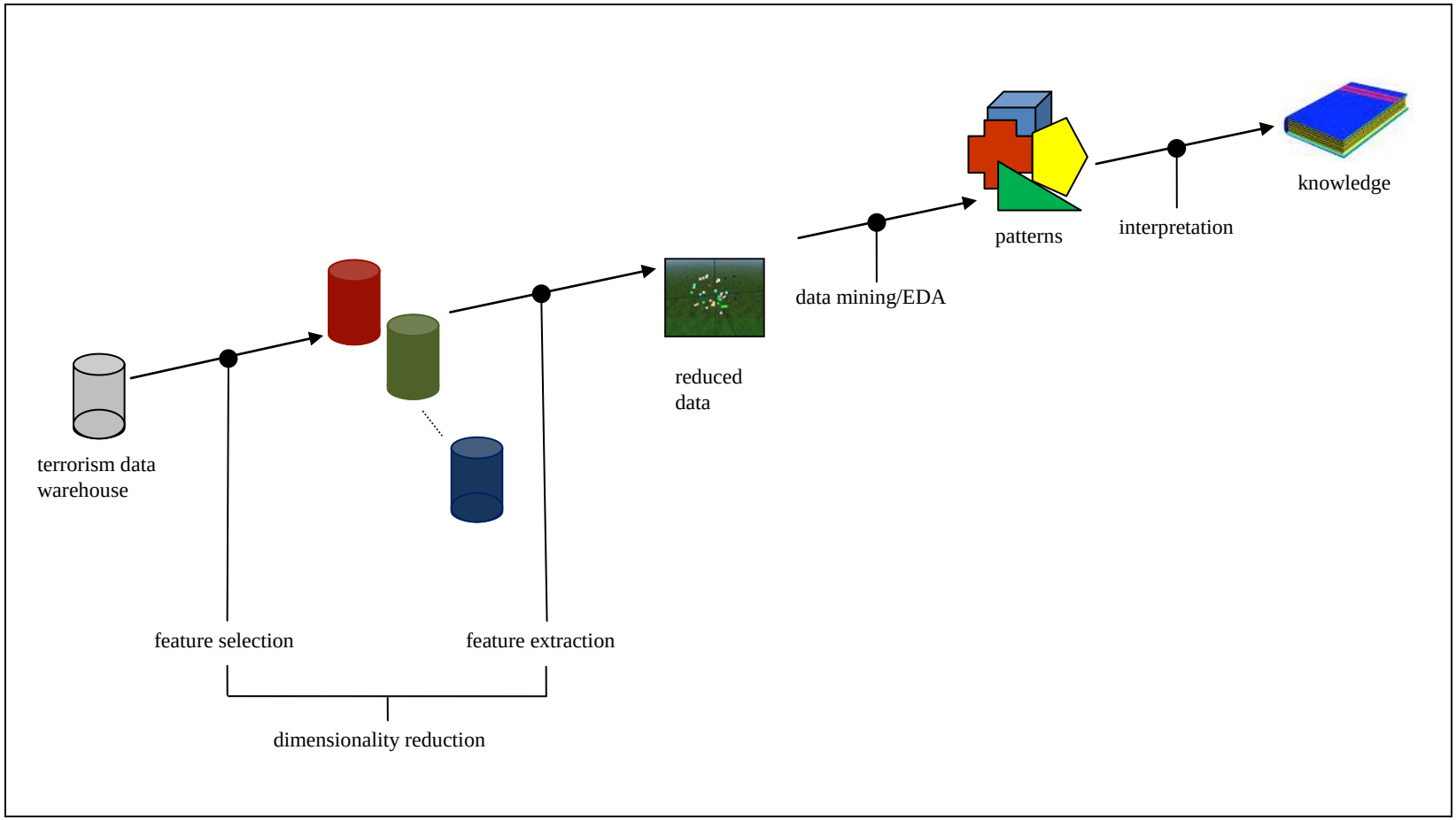

Figure 1. KDD process for terrorist data (adapted from Nieves and Cruz [*3*]).

2. Dimensionality Reduction for Data Visualization

Interpretation of any underlying structure for data in HDD space (d) is done by re-embedding into a lower 2-D/3-D Euclidean space. The projection could be for a nonlinear manifold, which is locally linear but may be globally curved. The projection should remain representative of the original data so that there is no loss of information and properties are preserved. DR of HDD is done here by FS and/or FE.

Note that DR tries to exploit the typically lower intrinsic dimension (P) of the data, i.e., $P < d$. P is the minimum number needed to account for observed properties of the data and reveals the presence of topological structure. Ideally, the reduced dimension (D) will correspond to P. When $P < D$, where D is also the dimension of the embedding space, then the data lies in a well-defined space.

2.1 Feature Selection

FS determines several features (or attributes) for the HDD by removing irrelevant and redundant data. An example of a decision system M, which can be represented as a matrix of objects and attributes, is illustrated in table 1 (*4*). The search for a feature subset involves determining those that are highly correlated with the decision attribute but uncorrelated with one another: or, in other words, compute the smallest subset of conditional attributes that preserve the decision attribute. This is called a *reduct*. The example shown in table 2 was computed using rough set theory (RST).* In this case, we obtained a 50% reduction of conditional data, i.e., we could safely eliminate half of the conditional attributes without changing the value of e, i.e., V_e.

Table 1. An example data table from Jensen and Shen (*4*).

$x \in U$	a	b	c	d	e
0	1	0	2	2	0
1	0	1	1	1	2
2	2	0	0	1	1
3	1	1	0	2	2
4	1	0	2	0	1
5	2	2	0	1	1
6	2	1	1	1	2
7	0	1	1	0	1

Note: For objects 0 to 7, the conditional attributes are from a to d, and the decision attribute is e.

*Only the results are shown here. A detailed description for this example is given in Jensen and Shen (*4*).

Table 2. Reduced data set for table 1.

$x \in U$	b	d	e
0	0	2	0
1	1	1	2
2	0	1	1
3	1	2	2
4	0	0	1
5	2	1	1
6	1	1	2
7	1	0	1

RST is an extension of conventional set theory, thus is discrete-based. Uncertainty is "indiscernibility" for a rough set attribute reduction. However, a "vagueness" of feature data, i.e., real-valued attributes, is not modeled.

Fuzzy-rough set theory (FRST) handles both discrete and continuous data. The implementation of fuzzy-rough feature selection (FRFS) being used in our work was written for the University of Waikato (NZ) environment for knowledge analysis (WEKA). WEKA (5) is a popular open-source environment. In particular, we are using the Java jar for ant colony optimization (ACO), i.e., FRFS-ACO, for a search of the feature space. FRFS-ACO requires a graph representation in determining the reduct.

Ideally, FS will result in 2-D/3-D data for VA application. For more than three-component results, the human visual system/brain combination usually becomes less effective quite quickly. In this case, an FE is then applied.

2.2 Feature Extraction

FE irreversibly transforms data semantics, but the underlying topology of the structure, if any, is preserved and can be further examined. In topology, the concern is not the representation of an object (or structure) in space, but the connectivity, which must not be altered. In other words, twisting, deforming, and/or stretching are allowed, but no tearing. For example, a 2-D circle is topologically equivalent to an ellipse.

Many FEs have been developed over the years. In figure 2, the first two approximations—principal component analysis (PCA) and classical metric multidimensional scaling (CMDS)—are a linear DR (LDR). An LDR is based on a linear combination of the feature data. LDRs keep similar data points close together (distance-preserving) when mapping from d to D. However, they cannot find curved manifolds since they are based on a Euclidean distance.

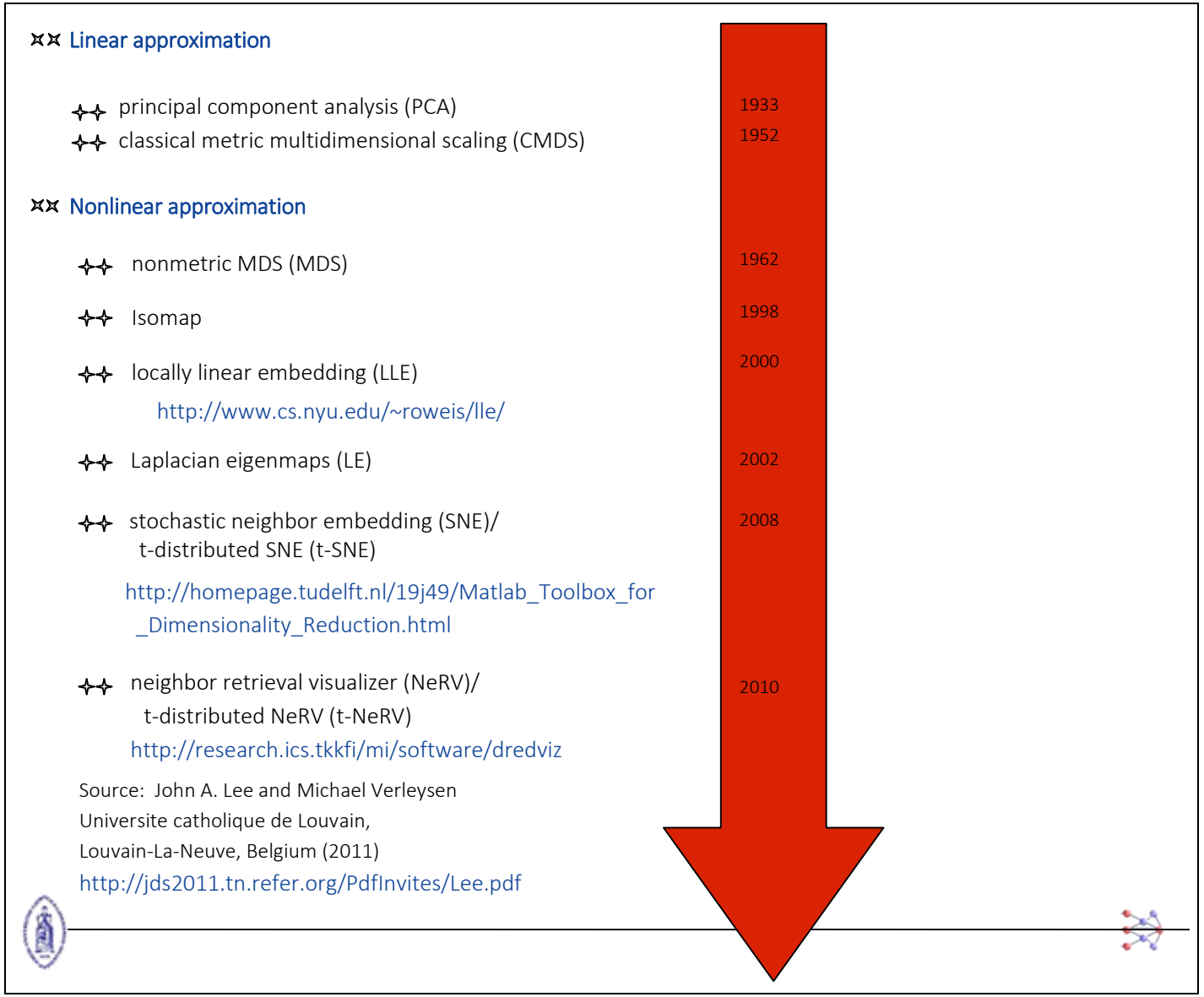

Figure 2. Some FEs for HDD and timeline.

A nonlinear DR (NLDR) approximation, which is also called a manifold learner, preserves geodesic distances along the manifold, linear or nonlinear (see figure 3 for a comparison between Euclidean, geodesic distance). NLDRs include nonmetric MDS, Isomap, LLE, LE, SNE/t-SNE, and NeRV/t-NeRV. Most papers for an NLDR approximation demonstrate the algorithm using an artificial dataset, such as the Swiss roll or S-curve, and thus are not straightforward to application of real-world data.

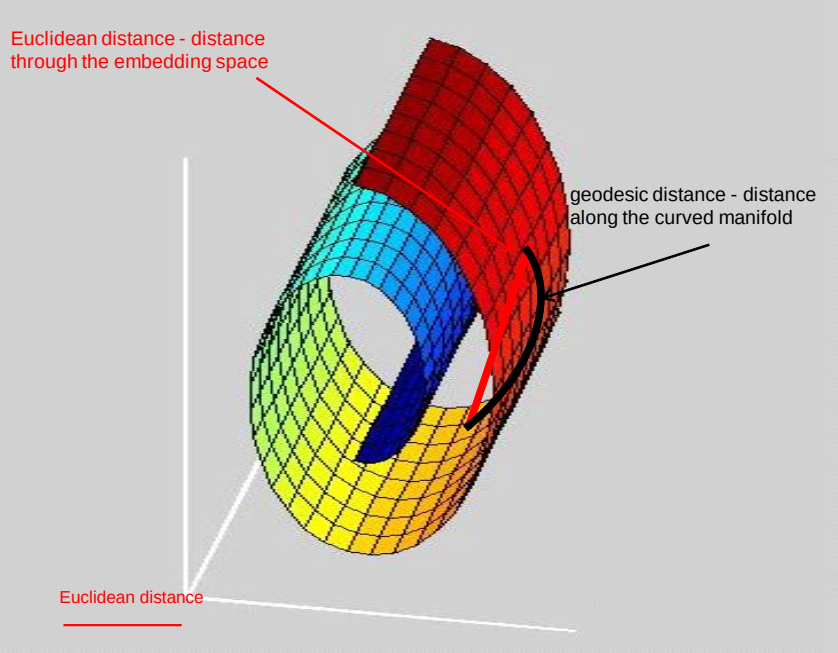

Figure 3. Comparison of Euclidean vs. geodesic distance. LDRs use metrics based on the Euclidean distance between two points, while the NLDRs are based on geodesic distance. An NLDR successfully unrolls the curved manifold, whereas an LDR fails.

A recent research paper (*6*) suggests that manifold learners may not be the best DRs for data visualization. The last two methods for NLDR in figure 2, namely SNE/t-SNE and NeRV/t-NeRV, are NLDRs specifically designed for data visualization, and have been used with real-world data; NeRV is an MDS for detecting local structures, i.e., an LMDS. That paper also states that SNE is a special case of NeRV ($\lambda = 1$ in equation 1 of the paper).

3. Exploratory Visual Analytics

As mentioned in the previous section, EDA in HDD space is done statistically using WEKA (figure 4). Launching the Explorer application from the graphical user interface (GUI) provides for FRST attribute reduction of HDD—specifically, an ant colony optimization (FRFS-ACO) search for reduct as described by Jensen (*7*). For a reduct that is >3, we then apply the neighbor retrieval visualizer (NeRV) (*6*).

Figure 4. WEKA GUI for data mining HDD using FRFS-ACO.

NeRV is a local MDS. Although semantics are destroyed by a feature extraction, the topology of the structure (or the random, scattered points) can then be inspected. Remember that the intent is to visually examine the data in a Euclidean space, i.e., EVA.

The resulting scene is described declaratively for the Extensible 3D (X3D) application-programming interface (API). X3D is an International Standards Organization (ISO) specification for describing scene content, possibly distributed across the Web. The scene graph consists of a directed acyclic graph of X3D objects and has a hierarchical parent-child structure. In addition, the immersive profile for the X3D scene allows for navigation/interaction within the

data. Details of all X3D nodes and attributes can be found at http://www.web3d.org /x3d/specifications/ISO-IEC-19775-X3DAbstarcrSpecification/; an excellent description of X3D nodes and concepts is also done by Brutzman and Daly (*8*).

In 2010, X3D nodes were tightly coupled with the HTML document object model (DOM) tree (*9*) for some Web browsers, such as Mozilla Firefox and Google Chrome. The result was an X3DOM library where one could embed X3D models directly into a Web page without having to write any JavaScript code. X3DOM uses the WebGL API to render interactive 3-D scenes natively in the Web browser.

4. Conclusions and Future Work

EVA in an HDD space for a timely interpretation remains to this day a very challenging task, especially for terrorist-related data. Dr. Nam suggests in her dissertation (*10*) that our perception in 3-D is learned from infancy, and that it is essentially nonexistent for higher dimensions. Thus it becomes more difficult in time to reason in higher dimensions.

Both feature selection and feature extraction, if necessary, are used for dimensionality reduction of HDD for data visualization. Declarative X3D and X3DOM are then used for VA of resultant data in either the latest Mozilla Firefox or Google Chrome Web browser; these also support WebGL for bringing 3-D to the Web browser procedurally.

Point characterizations constructed from 2-D orthogonal views of HDD, i.e., scatter plot diagnostics (scagnostics) and scatter plot matrix, are being considered (*11*). This approach to VA of HDD is guided by a more vigorous statistical analysis.

5. References

1. The University of Maryland's Global Terrorism Database Web site. National Consortium for the Study of Terrorism and Responses to Terrorism (START). www.start.umd.edu/gtd/ (accessed June 2012).
2. Lee, J. A.; Verleysen, M. *Nonlinear Dimensionality Reduction*; Springer Science+Business Media: New York, 2007.
3. Nieves, S.; Cruz, A. Finding Patterns of Terrorist Groups in Iraq: A Knowledge Discovery Analysis; *Proceedings of the 9th Latin American and Caribbean Conference for Engineering and Technology*, 3–5 August 2011; pp WE1-1– WE1-10.
4. Jensen, R.; Shen, Q. *Computational Intelligence and Feature Selection*; John Wiley and Sons, Inc.: Hoboken, NJ, 2008.
5. The University of Waikato Web site. http://www.cs.waikato.ac.nz/~ml/weka (accessed September 2012).
6. Kaski, S.; Peltonen, J. Dimensionality Reduction for Data Visualization. *IEEE Signal Processing Magazine* **2011**, 100.
7. Jensen, R. Performing Feature Selection With ACO. http://cadair.aber.ac.uk/dspace/bitstream/handle/2160/488/SIDM-03.pdf (accessed July 2012).
8. Brutzman, D.; Daly, L. *X3D: Extensible 3D Graphics for Web Authors*; Morgan Kaufmann Publishers Elsevier, Inc.: New York, 2007.
9. Behr, J.; Eschler, P.; Jung, Y.; Zollner, M. X3DOM – A DOM-based HTML/X3D Integration Model, 2009. http://www.x3dom/paper/2009-x3dom-web3d.pdf (accessed October 2012).
10. Nam, E. J. Exploratory Visual Analytics in High Dimensional Space. Ph.D. Dissertation, Stony Brook University, Stony Brook, NY, August 2011.
11. Wilkinson, L.; Anand, A.; Grossman, R. High-Dimensional Visual Analytics: Interactive Exploration Guided by Pairwise Views of Point Distributions. *IEEE Transactions on Visualization and Computer Graphics* **2006**, *12* (6), 1363–1372.

List of Symbols, Abbreviations, and Acronyms

API	application-programming interface
DOM	document object model
DR	dimensionality reduction
EDA	exploratory data analytics
EVA	exploratory visual analytics
FE	feature extraction
FRFS-ACO	fuzzy-rough feature selection using ant colony optimization
FRST	fuzzy-rough set theory
FS	feature selection
HDD	high-dimensional data
ISO	International Standards Organization
KDD	knowledge discovery in databases
LDR	linear dimensionality reduction
NeRV	neighbor retrieval visualize
NLDR	nonlinear dimensionality reduction
VA	visual analytics
WEKA	University of Waikato (NZ) environment for knowledge analysis
X3D	Extensible 3D

NO. OF
COPIES ORGANIZATION

1 (PDF) DEFENSE TECHNICAL INFORMATION CTR
DTIC OCA
8725 JOHN J KINGMAN RD
STE 0944
FORT BELVOIR VA 22060-6218

1 (PDF) DIRECTOR
US ARMY RESEARCH LAB
RDRL CIO LL
2800 POWDER MILL RD
ADELPHI MD 20783-1197

ABERDEEN PROVING GROUND

1 (PDF) DIR USARL
RDRL CII C
A NEIDERER

INTENTIONALLY LEFT BLANK.